State of OSG

Frank Würthwein
OSG Executive Director
UCSD/SDSC

March 14th 2022

Welcome !!!

SDSC SAN DIEGO SUPERCOMPUTER CENTER

We follow the OSG Code of Conduct :

https://opensciencegrid.org/management/conduct/

The OSG believes that a culture of inclusion, integrity, and cooperation is necessary to achieve its scientific goals, by affording all members the opportunity to reach their full potential. All OSG members, participants, vendors, staff, volunteers, conference and workshop attendees, invited speakers, and all other stakeholders are expected to conduct themselves in a professional manner that is welcoming to all and free from any form of discrimination, harassment, or retaliation. OSG members pledge to treat each other with respect and strive to model behaviors that encourage productive debate and allow for respectful disagreement.

All participants in OSG activities will not engage in any inappropriate actions or statements that are derogatory or defamatory on the basis of individual characteristics such as age, race, ethnicity, sexual orientation, gender identity, gender expression, marital status, nationality, political affiliation, ability, status, educational background and/or socioeconomic background, neurodiversity, mental or physical health, or any characteristics protected by law. Disruptive or harassing behavior of any kind will not be tolerated. Harassment includes but is not limited to inappropriate or intimidating behavior and language, unwelcoming jokes or comments, offensive images, photography without permission, and stalking.

Private chats are enabled throughout in zoom.
Please follow code of conduct, and especially in private chats.

Participants in OSG activities are encouraged to resolve any perceived breach of respectful decorum in a professional and informal manner before escalation. If an individual does not feel comfortable confronting the violation and/or believes someone has violated the code of conduct and it has not been addressed, they should report it by emailing conduct@opensciencegrid.org or discussing it with one of the standing OSG CoC appointees[1] who will follow-up on the reported violation in a confidential manner. The appointees will determine ways of redressing the matter and counsel the parties involved. Sanctions may be issued and range from verbal warning, ejection from a meeting without refund, removal of subscription from a forum or mailing list, revocation of access to OSG services, up to notifying appropriate authorities. Retaliation against the CoC appointees or the individual(s) reporting inappropriate conduct will not be tolerated. Appeals of sanctions for off-meeting violations, with long term impacts, may be directed to the OSG Council co-chairs.

1. At the moment, the OSG CoC appointees are the members of the OSG Executive Team.

Simple Rules for Virtual Meetings

- During sessions, all attendees are muted by default.
  - Only the Host & co-hosts can unmute you.
  - During the breaks, you will be able to unmute yourself, and are welcome to talk with others as you see fit.
- Raise your hand if you want to speak.
  - Co-hosts will call on raised hands during Q&A after each talk.
- Feel free to add questions and/or comments to the chat at any time.
  - Co-hosts will answer Q's and/or ask speaker during Q&A after talk.
  - If time runs out, speakers may answer Q's to their talks during the following talks in the same session.
- We will keep zoom session alive during breaks, and you are welcome to continue Q&A then.

The success of the virtual AHM depends on all of us working together within the limitations of being virtual.

OSG "Statement of Purpose"

OSG is a consortium dedicated to the advancement of all of open science via the practice of distributed High Throughput Computing (dHTC), and the advancement of its state of the art.

SDSC SAN DIEGO SUPERCOMPUTER CENTER

Four categories of participants in the OSG Consortium

- The individual researchers and small groups through the **Open Science Pool**.
- The campus Research Support Organizations
  - Teach IT/CI organizations & support services so they can integrate with OSG
  - Train the Trainers (to support their researchers)
- Multi-institutional Science Teams
  - XENON, GlueX, SPT, Simons, and many many more
  - Collaborations between multiple campuses
- The 4 "big science" projects:
  - US-ATLAS, US-CMS, LIGO, IceCube

OSG Vision & Aspiration

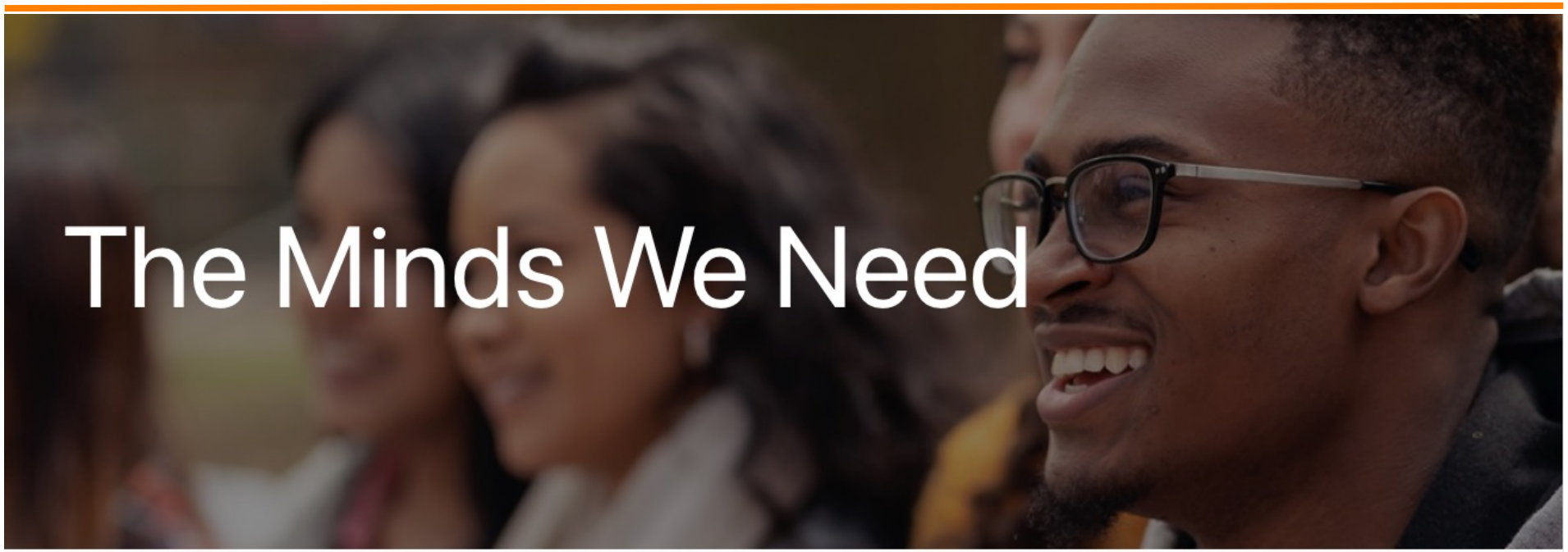
SDSC SAN DIEGO SUPERCOMPUTER CENTER

Long Term Vision

- Create an Open National Cyberinfrastructure that allows the federation of CI at all ~4,000 accredited, degree granting higher education institutions, non-profit research institutions, and national laboratories.
  - Open Science
  - Open Data
  - Open Source
  - Open Infrastructure
    - Open Compute
    - Open Storage & CDN
    - Open devices/instruments/IoT, …?

Openness for an Open Society

Democratizing Access



- **Connect every community college, every minority serving institution, and every college and university, including all urban, rural, and tribal institutions** to a world-class and secure R&E infrastructure, with particular attention to institutions that have been chronically underserved;
- **Engage and empower every student and researcher** everywhere with the opportunity to join collaborative environments of the future, because we cannot know where the next Edison, Carver, Curie, McClintock, Einstein, or Katherine Johnson will come from; and

(See Tuesday AM Session for more)

https://mindsweneed.org

OSG Compute Federation (OSCF)

149 "green dots" listed on this map

Compute Resources at 64 institutions in the USA alone.

(another 16 internationally that support science other than LHC via OSG)

Institutions contributing to OSCF

- **64 US institutions contributed compute resources** during the year ending on March 9th 2022.
  - We count institutions, organizations inside institutions & “clusters”
    - E.g. UCSD is an institution with 2 ”green dots”: SDSC & CMS group in Physics. 6 clusters contributed resources in 2021, incl. one entirely inside the commercial cloud, and another that itself has hardware across 30 institutions.
- Out of these 64 institutions in the USA
  - **9 are Minority Serving Institutions (MSI)**
    - 1 CC*, 2 EPSCoR, 1 Non-R1
  - 10 are in EPSCoR states
  - 17 received a CC* award for a compute cluster
  - 13 are non-R1
  - 31 are none of the above

26 of the 64 US institutions are either MSI, EPSCoR, or non-R1

OSG is Democratizing Access to Cyberinfrastructure

OSG Data Federation (OSDF)

20 Caches … 6 of which are in R&E network backbone

10 Data Origins … incl. one for the Open Science Pool

Federation = distributed control

- Resource owners determine policy of use for what they own.
  - Implies policy is set locally at the resource.
- Resource consumers determine policy of what they are willing to use.
  - Implies policy is set locally at the user access point.
- Federated system matches consumers to owners respecting both sets of policies.

The Power of Sharing

- Any participating institution shall be able to (dynamically) share any fraction of its resources with any other.
  - Collaborating Researchers can pool resources
- Institutions can share resources for the common good of all.
  - To democratize access funders can stipulate nationwide sharing for some fraction of the resources they fund.

Brief Overview of Concepts

Researcher connects to
compute & data resources
via an access point

Access Point

Institutions connect
compute resources via a
hosted compute entrypoint (CE)

Compute resources
are aggregated
in **compute resource pools**

jobs that do science
are submitted by researchers
at an access point
to a resource pool
to execute on compute resources
and access data via the **caches**
in the data federation

Institutions connect
data resources
via a **data origin**

(see Tuesday PM talk by Lauren & fkw for more on how this works)

The Open Science Pool Community

A community for all researchers in the US, from undergraduates to post-graduates.

We often abbreviate Open Science Pool to OSPool

We sometimes use OSPool as synonym for its community of users.

A Feature-Complete dHTC Environment

- **Open Science Pool**
  - Submission infrastructure that functions as compute “Access Point”
  - Workload management system
    - Complex workflows across heterogenous resources possible.
      - Easy to run workflows comprised of 100,000 jobs or more with complex dependencies between sets of jobs (full support of arbitrary DAGs).
  - Homogeneous runtime environment across heterogeneous resources
    - Includes a dozen or more types of GPUs
    - 100’s of curated containers
- **Open Science Data**
  - Storage that functions as “Data Origins”
  - Transparent “Data Access” via caches in the data federation
  - Supporting Public and Private Data

Any Researcher in the US can request access to this dHTC environment

Examples for Science on OSG

David Swanson Award Recipient:

Investigating the Strong Nuclear Force with the OSG (2022 David Swanson Awardee)	*Connor Natzke*
	11:30 - 12:00

Monday Afternoon Session:

Genetic Algorithm for Crystal Structure Prediction, using the computational resources of the OSG	*Kristal Varela*
	13:00 - 13:20
Computational Challenges and Opportunities in Multi-messenger heavy-ion physics	*Chun Shen*
	13:20 - 13:40
The Role of OSG in Advancing Research in Population Genetics	*Parul Johri*
	13:40 - 14:00
Enabling OSG through a student project to modify Airavata Science Gateway	*Derek Weitzel et al.*
	14:00 - 14:30

Tuesday PM for campuses & Thursday AM for science collaborations

The OSG Data Federation

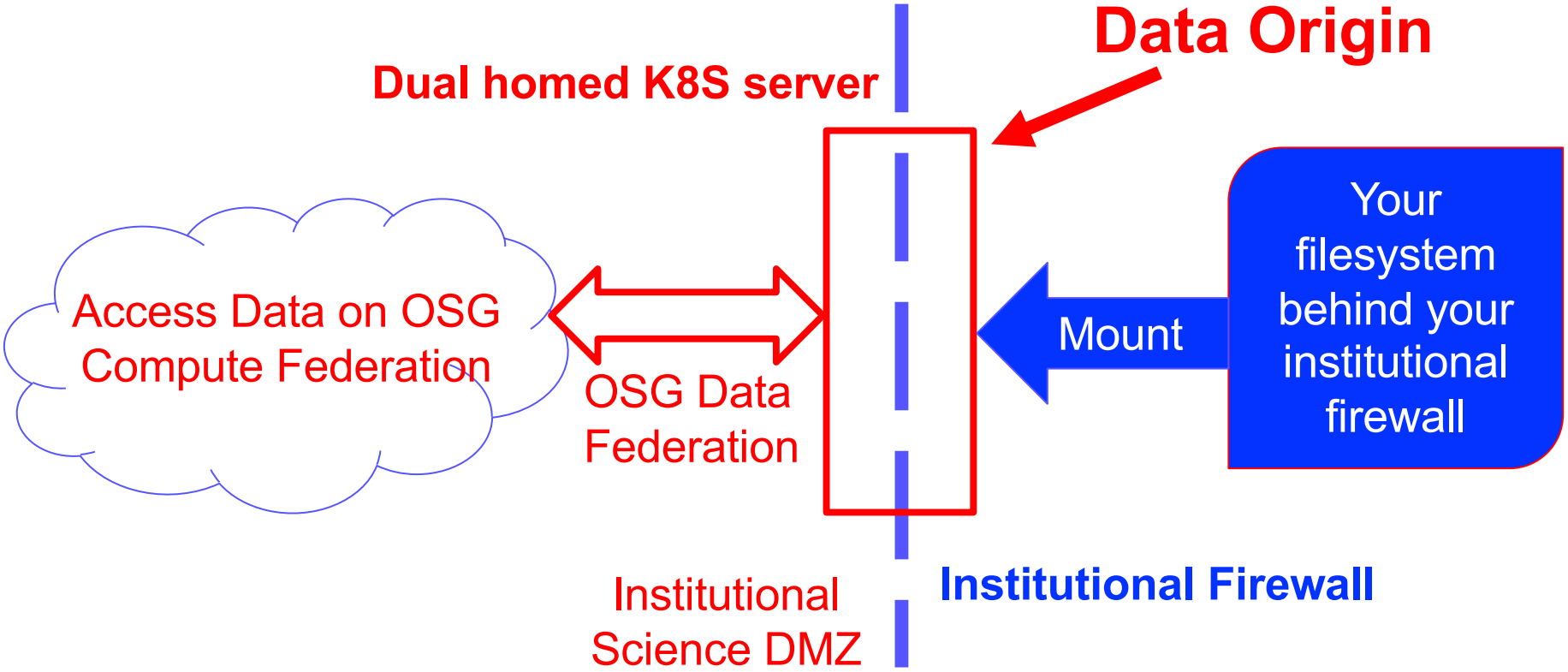
In 2021,
92 researchers, 9 collaborations, 1 campus
read 32PB of data out of a working set of 420TB
for an average re-read factor of 75.

You Can Join your data in 2 ways

1) Transfer your data via HTTPS to the PATh supported Origin
2) **Federate the filesystem at your institution with the OSDF**

SDSC SAN DIEGO SUPERCOMPUTER CENTER

Federating Data with the OSDF



Your fileserver with your data can be behind your institutional firewall.
A dual homed K8S server mounts only those filesystems you want to export.
We operate the OSDF API by deploying a “Data Origin” container into K8S.

(more on joining OSDF and OSCF in Tuesday PM Session)

Functionality of a Data Origin

- Export your data read-only into the Data Federation
  - You choose what part of your filesystems namespace you want to export.
  - You can change this dynamically any time you want.
  - **Data can be public or private**
  - Origin uses HTTPS as protocol => works as general webserver in addition to OSG Data Federation.
- Store output data produced on OSG
  - Put via HTTPS as part of HTCondor workflows
    - authorized only to those people you want to support.
  - Read-only access possible to data stored this way.
    - What is put into an origin may be read via the federation if you so desire.

OSDF's Global Namespace

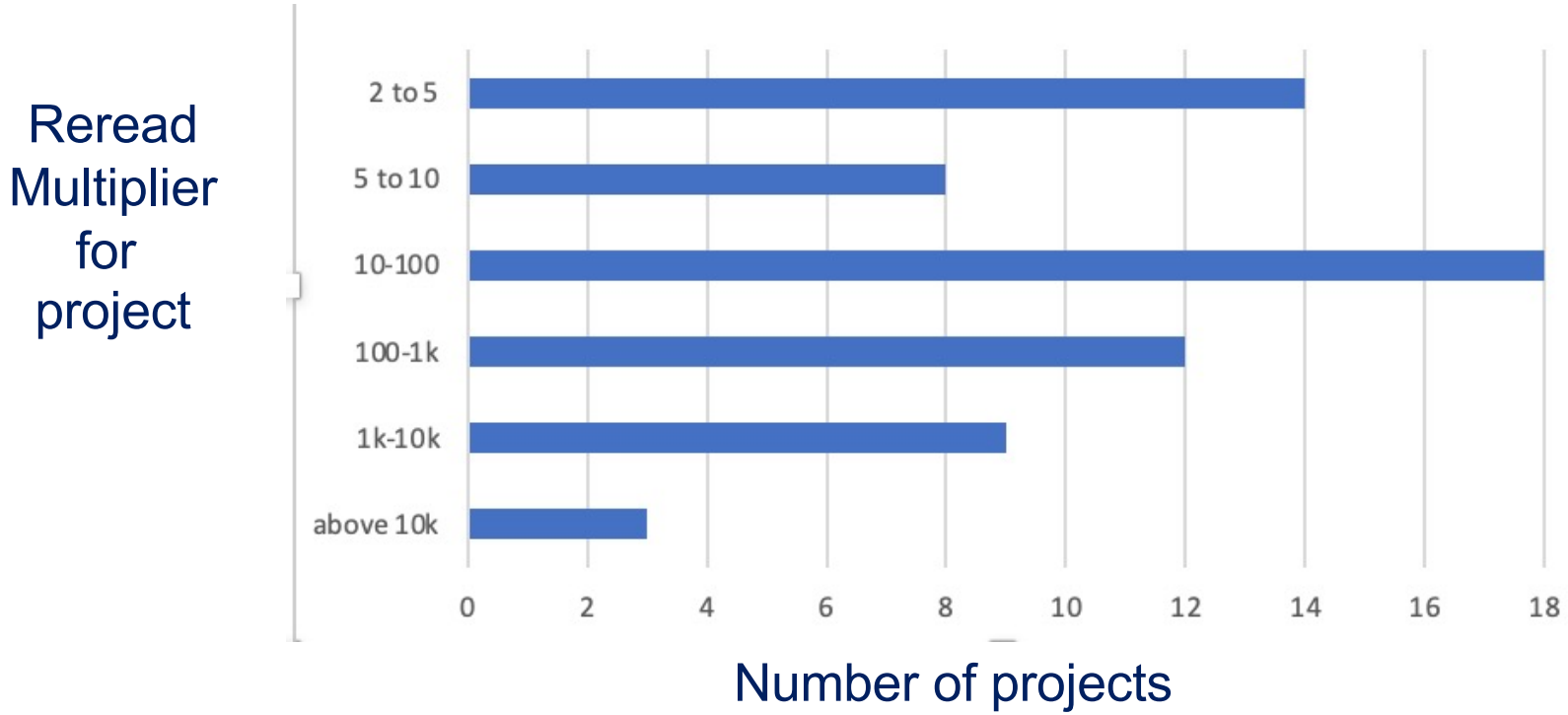
- Global Namespace is separate from the origins that hold the data
  - **You can move data between origins via HTTPS without changing how the data is accessed via the OSDF.**
    - Literally, nobody will notice !!!
- This allows federation of namespace that is separate from federation of server hardware that serves the namespace.
  - Lot's of interesting ways of using the power this provides you with.

Relationship to Access Point (AP)

- All data in the federation is visible to those authorized to see it from the AP.
  - AP has same access as runtime environment.
- Any data visible to the AP can be made visible to the runtime environment via the data federation.
  - Typically, you don't want anything in your home directory visible in the federation. So defaults are such that "home" is not visible.

Use of Data Federation in 2021

92 researchers, 9 collaborations & 1 campus read 32PB of data out of a working set of 420TB for an average re-read factor of 75.



Top Users of Data Federation

Project	Data Read	Working Set	Reread multiplier
LIGO (Private)	10PB	38TB	264
Minerva	5.6PB	3.1TB	1,789
NOVA	2.6PB	1.9TB	1,348
LIGO (Public)	2.4PB	38TB	67
Tufts_Hempstead	2.0PB	380GB	5,321
DUNE	1.6PB	185GB	8,658
Steward	1.0PB	11TB	92
REDTOP	874TB	95TB	9.2
Molcryst	570TB	5GB	115,650
BiomedInfo	530TB	66GB	8,090

17 projects have >TB working sets
11 of these are OSPool users

Tufts Computer Architecture Lab

Steward Observatory Data Analytics

R&D towards future particle physics experiment

Quantum chemical and machine learning insights into supra-molecular organization of molecular crystals

Development and application of software tools for performing large-scale biomedical informatics on microbial genome sequence data.

Advancing the State of the Art

This year's biggest stories of change:

- PATh Credit & PATh facility (see previous & next talks)
- Replacing authorization based on person with authorization based on capability.
  - The "Token" Transition

See Wednesday Afternoon Session:

13:00	**OSG 3.6 and Token Transition Update**	*Brian Lin* 13:00 - 13:25
	Handling HTC Jobs Tokens with Vault	*Dave Dykstra* 13:25 - 13:45
	Updates on the OSDF Monitoring System	*Derek Weitzel* 13:45 - 14:05
14:00	**Kubernetes at UChicago: PATh, IRIS-HEP, and the ATLAS AF**	*Lincoln Bryant* 14:05 - 14:30

Summary & Conclusion

- OSG continues to **advance all of open science via the practice of dHTC, and the advancement of its state of the art**.
  - Lot's of "Big Data" across many science domains
- Open Science Pool as strategy to **democratize access to dHTC**

Acknowledgements

- This work was partially supported by the NSF grants OAC-2030508, OAC-1841530, OAC-1836650, and MPS-1148698